BIOINFORMATICS APPLICATIONS NOTE

Vol. 28 no. 4 2012, pages 601–602
doi:10.1093/bioinformatics/btr705

Databases and ontologies

Advance Access publication December 30, 2011

SNPdbe: constructing an nsSNP functional impacts database

Christian Schaefer[1,2,*], Alice Meier[1], Burkhard Rost[1,2] and Yana Bromberg[3]

[1]Technische Universitaet Muenchen, Bioinformatics – I12, Informatik, Boltzmannstrasse 3, [2]Technische Universitaet Muenchen Graduate School of Information Science in Health (GSISH), Boltzmannstrasse 11, 85748 Garching, Germany and [3]Department of Biochemistry and Microbiology, School of Environmental and Biological Sciences, Rutgers University, New Brunswick, NJ 08901, USA

Associate Editor: Jonathan Wren

ABSTRACT

Summary: Many existing databases annotate experimentally characterized single nucleotide polymorphisms (SNPs). Each non-synonymous SNP (nsSNP) changes one amino acid in the gene product (single amino acid substitution;SAAS). This change can either affect protein function or be neutral in that respect. Most polymorphisms lack experimental annotation of their functional impact. Here, we introduce SNPdbe—SNP database of effects, with predictions of computationally annotated functional impacts of SNPs. Database entries represent nsSNPs in dbSNP and 1000 Genomes collection, as well as variants from UniProt and PMD. SAASs come from >2600 organisms; 'human' being the most prevalent. The impact of each SAAS on protein function is predicted using the SNAP and SIFT algorithms and augmented with experimentally derived function/structure information and disease associations from PMD, OMIM and UniProt. SNPdbe is consistently updated and easily augmented with new sources of information. The database is available as an MySQL dump and via a web front end that allows searches with any combination of organism names, sequences and mutation IDs.

Availability: http://www.rostlab.org/services/snpdbe

Contact: schaefer@rostlab.org; snpdbe@rostlab.org

Received on September 9, 2011; revised on November 11, 2011; accepted on December 17, 2011

1 INTRODUCTION

Resources like dbSNP (Sherry *et al*., 2001) and UniProt (Bairoch *et al*., 2005) contain many experimentally determined nsSNPs, but few of these are annotated with respect to function. Some databases [e.g. PMD (Kawabata *et al*., 1999)] contain experimental annotations of functional effects of mutants. However, these are sparsely populated and do not directly link to dbSNP or UniProt. For the vast majority of mutations lacking experimental annotation, we can gauge functional impact only via *in silico* analysis.

Proper use of computational methods requires specific skills and resources generally inaccessible to medical researchers or experimental biologists. To help, we created an MySQL database readily usable by non-experts. We collected SAASs from PMD, dbSNP, 1000 Genomes (1000_Genomes_Project_Consortium, 2010) and UniProt 'variant's and 'mutant's. We also store 'conflict' records to illustrate how sequencing discrepancies may lead to differing interpretations of the functional significance of a given sequence position. For each SAAS we predict the functional effect using SNAP (Bromberg and Rost, 2007) and SIFT (Ng and Henikoff, 2001). Where available, predictions are augmented by experimental annotations and associated human diseases. We also compute evolutionary conservation of the mutant positions. A web interface provides convenient access to underlying data via organism, sequence and mutation ID queries.

*To whom correspondence should be addressed.

2 DATA SETUP AND RETRIEVAL

Database: SNPdbe mutation data comes from dbSNP, UniProt, 1KG and PMD (Fig. 1A). UniProt and PMD store protein sequences explicitly, while dbSNP links to RefSeq (Pruitt *et al*., 2007). dbSNP collects 1KG variants with a time delay, so for SNPdbe we mapped all 1KG nsSNPs to RefSeq using Annovar (Wang *et al*., 2010). We keep only one version of redundant protein sequences, referenced by md5 checksums irrespective of origin. Redundancy is assessed at full-sequence identity (maximum one substitution per sequence) over the entire sequence (+/− leading Met residue). This allows correlating mutations from different sources referencing the same sequence. We currently store 1 362 793 unique SAASs in 158 004 proteins from 2684 organisms covering all kingdoms of life; the top five contributors are human, mouse, rice, cow and rat. For each SAAS we provide the following information: (i) SNAP and SIFT binary predictions of functional effects (neutral/non-neutral). (ii) Evolutionary conservation information from PSIC (Sunyaev *et al*., 1999), PSI-Blast (Altschul *et al*., 1997) PSSMs and frequency scores from runs against PDB (Berman *et al*., 2000) and UniProt. (iii) Functional effects from PMD and UniProt. For human SAASs, disease associations are also available from PMD, UniProt and OMIM (Amberger *et al*., 2009) (Fig. 1B). (iv) dbSNP evidence and average heterozygosity, and (iv) interesting functional/structural features (UniProt) at the mutation site. Data are stored in an MySQL database and are downloadable as a dump file.

Web interface: The database is web-accessible allowing gene/protein ID/name, disease, sequence (or its md5 hash) and mutant-based queries. Some queries (e.g. md5, gene ID) are exact. Sequence queries are BLAST similarity based. Keyword searches (e.g. disease) are 'loose', i.e. matched to corresponding free text fields. The results page lists all SAASs found within the specified sequence and their functional effect predictions, wild-type/mutant conservation scores, information on disease (human only), experimentally derived functional/structural consequences, changes in position biochemical properties, per-variant validation status and average heterozygosity. This information is also

© The Author(s) 2012. Published by Oxford University Press.
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/3.0), which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

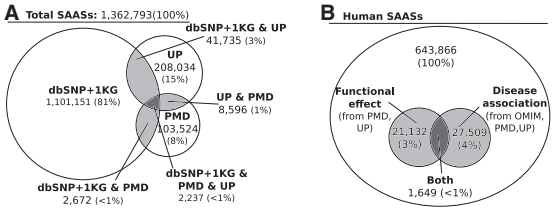

Fig. 1. Venn diagrams describing the overlap of (**A**) all SNPdbe component databases and (**B**) functional and disease annotations of human SAASs. Note that <1% of human SAASs have both functional effect and disease annotations.

accessible via single/batch mutation queries with dbSNP rsids, PMD or SwissVar IDs or SAASs in the X*pos*Y format (and associated sequence). The user can (i) restrict queries to specific organisms or protein keywords; and (ii) search for mutants in similar sequences. Query results may be sorted by different attributes and downloaded in CSV format. Linkouts to referenced web resources are available.

Example: dbSNP rsid 104894374 describes the mutation R157W in the RDH5 gene. This mutation is associated with eye disease, *Fundus albipunctatus* (OMIM 601617.0008). Both SNAP and SIFT predict this substitution to be non-neutral. Indeed, it results in loss of activity in the gene product (PMD A010122). By combining mutation disease associations and their functional effects new inferences can be made about molecular functions altered in disease.

3 CONCLUSION

SNPdbe is designed to fill the annotation gap left by the high cost of experimental testing for functional significance of protein variants. It joins related bits of knowledge, currently distributed throughout various databases, into a consistent, easily accessible and updatable resource. The major features distinguishing SNPdbe from other databases are: (i) the inclusion of a much wider array of organisms and data sources; and (ii) the explicit differentiation between functional/structural effects and disease associations. Furthermore, unlike SNPdbe, existing resources (i) lack experimental annotation of functional/structural changes or offer only single tool (e.g. SIFT) predictions (Mooney and Altman, 2003; Thorn *et al*., 2010), (ii) are limited to naturally occurring variants (Chelala *et al.*, 2009), (iii) are not consistently updated (Jegga *et al*., 2007; Wang *et al*., 2006) or (iv) do not offer pre-computed effects on a large scale (Reva *et al*., 2011; Wang *et al*., 2010). SNPdbe's database schema and management scripts are designed to easily handle the addition of new sequences and SAASs and the integration of new predictors and sources of experimental data. Monthly updates are planned. Information about current versions of included databases and statistics is available from SNPdbe website. Our ultimate goal is to make SNPdbe a toolbox for biologists and medical researchers dealing with mutation data. Computationally acquired predictions and annotations found in SNPdbe will help design and prioritize further experimental research.

ACKNOWLEDGEMENTS

We thank to Laszlo Kajan, Guy Yachdav and Tim Karl (TUM) for maintenance of our compute cluster and to those who deposit their experimental data in public databases.

Funding: Alexander von Humboldt Foundation (to C.S., A.M. and B.R.); Rutgers, New Brunswick, start-up funds (to Y.B.).

Conflict of Interest: none declared.

REFERENCES

1000_Genomes_Project_Consortium (2010) A map of human genome variation from population-scale sequencing. *Nature*, **467**, 1061–1073.

Altschul,S.F. *et al.* (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.

Amberger,J. *et al.* (2009) McKusick's Online Mendelian Inheritance in Man (OMIM). *Nucleic Acids Res.*, **37**, D793–D796.

Bairoch,A. *et al.* (2005) The Universal Protein Resource (UniProt). *Nucleic Acids Res.*, **33**, D154–D159.

Berman,H. *et al.* (2000) The Protein Data Bank. *Nucleic Acids Res.*, **28**, 235–242.

Bromberg,Y. and Rost,B. (2007) SNAP: predict effect of non-synonymous polymorphisms on function. *Nucleic Acids Res.*, **35**, 3823–3835.

Chelala,C. *et al*. (2009) SNPnexus: a web database for functional annotation of newly discovered and public domain single nucleotide polymorphisms. *Bioinformatics*, **25**, 655–661.

Jegga,A.G. *et al.* (2007) PolyDoms: a whole genome database for the identification of non-synonymous coding SNPs with the potential to impact disease. *Nucleic Acids Res.*, **35**, D700–D706.

Kawabata,T. *et al*. (1999) The Protein Mutant Database. *Nucleic Acids Res.*, **27**, 355–357.

Mooney,S.D. and Altman,R.B. (2003) MutDB: annotating human variation with functionally relevant data. *Bioinformatics*, **19**, 1858–1860.

Ng,P.C. and Henikoff,S. (2001) Predicting deleterious amino acid substitutions. *Genome Res.*, **11**, 863–874.

Pruitt,K.D. *et al*. (2007) NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res.*, **35**, D61–D65.

Reva,B. *et al*. (2011) Predicting the functional impact of protein mutations: application to cancer genomics. *Nucleic Acids Res.*, **39**, e118.

Sherry,S.T. *et al.* (2001) dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.*, **29**, 308–311.

Sunyaev,S.R. *et al.* (1999) PSIC: profile extraction from sequence alignments with position-specific counts of independent observations. *Protein Eng.*, **12**, 387–394.

Thorn,C.F. *et al*. (2010) Pharmacogenomics and bioinformatics: PharmGKB. *Pharmacogenomics*, **11**, 501–505.

Wang,K. *et al*. (2010) ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.*, **38**, e164.

Wang,P. *et al.* (2006) SNP Function Portal: a web database for exploring the function implication of SNP alleles. *Bioinformatics*, **22**, e523–e529.